METHODS OF PROCESSING TEXT FOUND IN IMAGES

The World Wide Web is a distributed database including hundreds of millions of documents. Search engines such as Alta Vista attempt to index the web based on ASCII text included on each page and on associated meta tags. Increasingly, however, text information is present on the Web in the form of text images. Known search engines are unable to make use of text presented in this form.

One approach to this problem is discussed in Lopresti et al, "Locating and Recognizing Text in WWW Images," Information Retrieval, vol.2, no.2-3 p.177-206, 2000, and involves a procedure based on clustering in color space followed by a connected-components analysis. Character recognition is performed using polynomial surface fitting and "fuzzy" n-tuple classifiers. While suitable for some applications, such techniques are too computationally intensive and imprecise for widespread use.

In accordance with one embodiment of the present invention, an image containing text is digitally watermarked with an identifier. The identifier serves as an index to a database record where additional information about the image, including keywords or full text of the included text, are provided. To obtain the associated data, a search engine web crawler or other process can download an image, apply a watermarking detection procedure, use an identifier thereby obtained to index a database, and access keywords or full text represented in the image from the indexed database record.

The text can be entered in the database using various known methods. One is to have the text manually coded by a clerical service. Another is to apply an automated OCR process to the image data, such as that detailed by Lopresti. Once the text is once thereby developed, it can be made quickly available repeatedly thereafter by reference to the associated database record.

The database can be conventional, and is preferably accessible over the internet. A suitable database system is disclosed in copending application 09/571,422, filed May 15, 2000. A variety of watermarking techniques are known. An illustrative set of techniques that can be employed in this application is disclosed in copending application

09/503,881, filed February 14, 2000. The disclosures of these applications are incorporated herein by reference.

The technology disclosed herein finds myriad applications. As noted, one is in the indexing of a collection of electronic documents (e.g., web pages). An index augmented by the results of such a procedure is generally more useful than such an index without augmentation.

Another application is in the use of webcams, or security monitoring cameras. Certain image frames from such sources (e.g., one every minute, or one every second, etc.) can be analyzed for textual information (e.g., license plate markings, superimposed date data), and the textual information stored. The image data is watermarked, with the watermark indicating the repository of the corresponding textual information.

Still another application is PDF documents or fax data files. (While some PDF files include corresponding ASCII text data, most do not.) The file data can be applied to an OCR engine, and the resulting text stored in a database. The PDF or fax data file can be slightly altered to impart a watermark – the watermark again serving to point to the repository of the corresponding text information.

Yet another application is in photocopiers. Again, the textual content is extracted from the scanned image of the original document. In this case the paper photocopy output (or a corresponding digital file) is altered in slight respects to encode a watermark. The watermark points to the text data repository.

While the illustrative embodiment particularly considered watermarks that convey an index to a remote database, other arrangements are naturally possible. For example, the watermark can directly encode the fulltext or keywords (forms of metadata).

Similarly, while the illustrative embodiment particularly considered imaged text in image files, the same principles can be applied more widely. For example, all metadata associated with an image through a watermark can be employed in compiling an index of the web or other collection of content data – not just included text (e.g., names of persons and places, dates, times, and other more application-specific metadata). Moreover, such techniques are not just limited to images. Other forms of content,

including video and audio, can be watermarked, and the metadata thereby associated with the content can be used for web indexing and other purposes.